智慧中医：肺癌处方智能生成模型

阮春阳 裴朝翰 张彦春 杨蕴 田建辉

关键词 智慧中医;深度学习

1 引言

作为人工智能化的支撑技术，深度学习已经在医学影像诊断、电子病历预测等医疗诊断任务中表现出色，发挥着重要作用。深度学习方法成熟之前，传统机器学习方法包括回归分析、支持向量机等是医学诊断任务的主要支撑方法，该类方法通过特征工程选取合适的特征值，在医学数据尤其是小数据集上表现出色。与此同时，随着中医信息化发展，传统机器学习方法在探寻中药药性和毒性等中医知识、促进中医诊断的标准化和客观化发挥了重要作用。与传统机器学习算法通常需要专家在原始数据上进行特征选择不同，深度学习算法作为多层次的表示学习算法，可以从原始数据中自动提取特征，并且逐层将低层次的表示抽象为更高层次的表示[1]。深度学习已经在拥有丰富数据集的西医临床诊断任务中得到了一定程度的有效应用，利于基于深度学习的新药发现[2]，电子病历诊断预测[3]等。 通过文献调研，深度学习模型在中医诊断中的应用尚处于起步状态，主要是通过深度卷积网络对中医临床舌诊，CT 图像数据进行分类[4]。如何基于深度学习模型构建智慧诊断模式并将其应用到实际中医诊断过程中，达到辅助临床的作用，是中医人工智能研究中一项重要课题。

中医临床诊断是一种个性化医疗模式。医生根据病人症状推断出病证，基于可能的病症医生做出疾病判别并给出相应的针对症状的中药。因此，中医诊断数据的各项特征十分复杂。例如，中药之间存在协同关系和多重共线性关系，中药和症状之间存在复杂的映射关系。图 1 展示了实际的中医临床数据和其中复杂关系模式。如何准确的去模拟这些实体间复杂关系，挖掘实际诊断模式给中医诊断智能化提出了挑战。针对以上挑战，本文探讨如何利用新兴深度学习模型构建智慧中医诊断新模式。

2 中医肺癌临床诊断

在我国恶性肿瘤发病率中，肺癌发病率居高不下。据国家癌症中心统计显示，2018 年我国新发肺癌病例约为 78.7 万例，发病率和死亡率分别达到 57.13/10 万，45.80/10 万人。在我国，中医药治疗恶性肿瘤历史悠久，拥有独特的理论体系和确切的治疗效果。特别是近半个世纪来，开展的多项中医药及中西医结合治疗肿瘤的临床研究，证实了其确切疗效[5]。中医诊断肺癌，首先通过“望”“闻”“问”“切”获取患者的症状和体征，根据症状判定患者的证型，最后根据病变的具体部位和证型选定一个主方，并在此基础上进行中药加减。

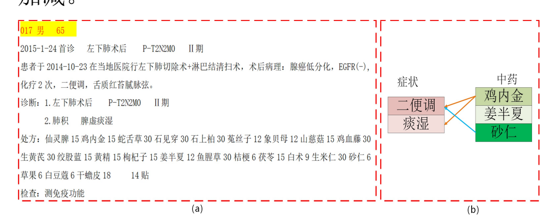


图 1 (a) 中医临床病历样例； (b) 中医病历复杂模式样例。

整个诊断过程会以文本形式记录并形成处方。现代医学诊断中，疾病拥有标准化的诊断和治疗规范，但中医对于患者的治疗更具有个体化特征，处方治疗效果与医生的临床水平具有强相关性。中医临床诊断既有背景知识的运用，逻辑规则的运用，也有大量不确定性问题的分析和求解，这正是医生辨证施治的精华[6]。基于以上中医肺癌临床诊断模式分析，如何利用处方数据挖掘其中用药等规律，构建智能处方生成系统辅助临床诊断，对中医传承与创新发展具有深远意义。目前，由于中医处方却少数字化和标准化，人工智能技术在其上的应用十分有限。本文在中医知识~~文本~~挖掘工作积累的基础上[7, 8, 9]，对中医肺癌临床处方数据调研分析，针对其数据特点，构建深度学习模型挖掘处方中症状和中药之间隐藏关系等规律，在此过程中与医生沟通验证模型的准确性，最终实现处方智能生成并达到较高的临床有效性，辅助医生诊断，提升临床效率，推动临床诊断创新发展。图 2 描述了本文整体模型构建思路。

3 基于循环神经网络的肺癌处方智能生成

循环神经网络 RNN 是一种特殊的神经网络结构，它是根据“人类认知是基于过往的经验和记忆”

这一观点提出的。不同于一般深度网络，RNN 不仅考虑前一时刻的输入,而且赋予了网络对前面输入内容的一种‘记忆'功能。RNN 之所以称为循环神经网络，是因为一个序列当前的输出与前面的输出也有关。

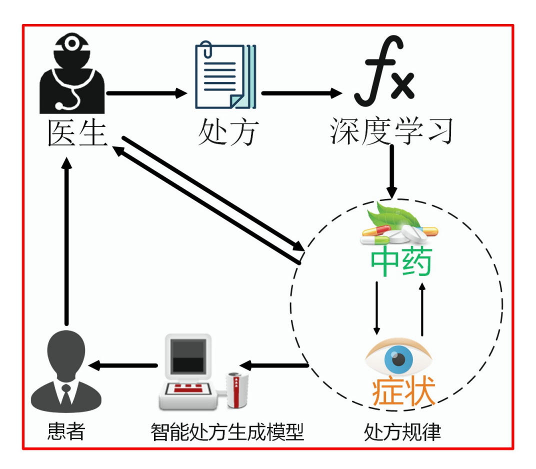


图 2 模型构建思路

具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。

本文将中医临床处方的生成过程建模为一个机器翻译的问题，输入的是一组症状序列，输出的是处方需要的草药组合，就像将一组由中文词汇组成的句子翻译成英文一样。而循环神经网络 RNN 非常适合处理这种序列与序列之间的翻译问题，它在自然语言处理的机器翻译方面有着广泛的运用，所以本文可以在中医的处方生成问题上借鉴已有的机器翻译的经验，来推陈出新，构建符合中医临床处方数据特点的智能处方生成系统。

编码器-解码器模型

针对将一个输入序列翻译成输出序列的问题，有人提出了编码器-解码器模型，所谓编码，就是将输入序列转化成一个固定长度的向量；解码，就是将之前生成的固定向量再转化成输出序列 [109]。

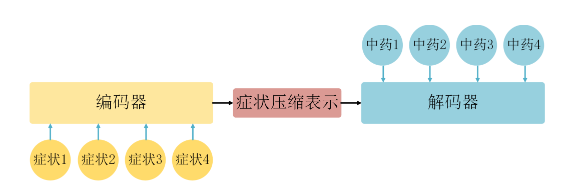


图 3 “ “症状-中药” ” 自编码-解码器模型

图 3 就描述了一个简单的自编码-解码模型器结构。利用自编码器，本文将一组中文词语序列，比如一组症状按照次序依次输入编码器，利用编码器将该组序列压缩成一个向量元素，及就是将一组症状聚合，得到症状组的压缩表示，这种压缩的症状表示是提取了症状序列中关键的特征信息。随后，本文将得到输入序列的压缩表示输入到解码器中，基于处方数据标签，解码器会将本文需要的输出序列元素，也就是草药依次解码出来。到此，本文给出了处方生成的基本思路和模型框架。但是，编码器-解码器仅仅是一个抽象的模型框架，它们内部的具体实现可以有许多种，但是在本文的中医处方生成任务中，比较适合的还是前面提到的循环神经网络 RNN。因为 RNN 具有的记忆功能，这一特性使编码器在一个一个输入症状，对齐进行压缩的时候，输出的中间压缩向量表示可以对前面输入的症状有“记忆”。在输出的时候，按照它所“记忆”的内容进行针对性的解码出需要的中药。

注意力（attention）机制

在实际模型运行中，基于 RNN 的自编码器处理一般文本序列数据已经取得了明显效果，但中医处方数据不同于一般的中文数据有序的组织形式，处方中症状、中药完全是无序的组织。基于 RNN 的自编码器在生成处方过程中会表现出了不足。如果利用循环神经网络 RNN 作为其基本组件，虽然 RNN 有一定的记忆能力，但是它还是随着距离衰减。拿中医症状来说，本文得到的症状压缩表示是在输入最后一个症状之后才生成的，所以它必然对离她比较近的症状“记忆”的比较多，离它远的，比如第一个输入的症状，压缩表示所“记忆”的信息就比较少了。

针对这种问题，本文采用用 RNN 的一些修正模型比如长短时记忆网络（LSTM），门控循环单元（GRU），代替简单 RNN 作为基本的编码器-解码器组件。利用 RNN 的改进形式虽然可以一定程度上解决在编码过程中的“遗忘”问题。但是就这个编码-解码过程而言，有一个基本的逻辑缺陷。

本文在将一组中医症状压缩表示之后，在利用其解码出来草药的过程中，每次解码利用的都是同一个压缩表示，但是这样是有悖于正常的思考模式的。举例来说，医生在开出某种中药的时候，不可能平均地考虑所有的症状，必然是有所侧重于某些症状，也就是说会专门地注意某种或某些

症状而不是对所有的症状分配相同的注意力。依据这种思路，就自然地引入了注意力（attention）机制[110]。

在本文之前的模型里，症状的压缩表示，在编码过程结束之后就再也不会改变了，也就是说本文在解码中药过程中，每次对于之前症状的考虑，或者是分配的注意力是相同的，这种情况明显是不现实的。本文的注意力机制，来源于一种对人类思考模式的模仿。在注意力机制的实现层面，本文在解码的过程中将不变的总体症状压缩表示，替换成可变的每个症状压缩表示的加权和。在循环神经网络里，每输入一个症状其实就可以产生一个压缩表示，它表示的是包含这个症状以及前面距离比较近的症状的大部分信息，所以本文没有必要采用等所有症状都输入完成之后所得到的压缩，而利用在输入每个症状时产生的压缩。他们带有一种局部性的信息，本文可以将他们加权求和，在本文需要重点考虑某些局部症状的时候，这些局部症状对应的压缩的权重就可以适当上调。

总之，本文解码不同中药所依据的症状压缩不再都是相同的一个，而是根据不同中药的情况，由每个症状的压缩在乘上实际考虑的不同权重求和而得的新的症状压缩表示。这就是本文所说的注意力机制。图 4 是加入注意力机制的自编码器模型。

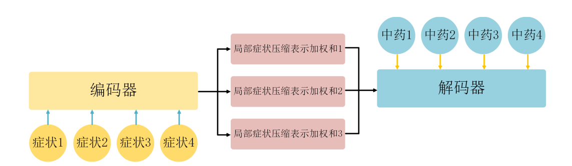


图 4 “ “症状-中药” ” 注意力机制自编码器模型

另外，在本文获取每个症状的局部压缩表示的时候，仅仅是获取了这个症状之前包含这个症状的信息，而不包含之后输入症状的信息。所以本文可以把输入症状的顺序变换一下，从后往前输入，这样本文就可以获得每个症状及其按照原始顺序来看之后的症状的信息，本文将其称之为反向的压缩表示。本文将正向和反向的压缩表示结合起来，可以代表某个输入症状前面和后面一段距离的局部情况，这比本文前面提到的单向的方式能包含更多的信息。图 5 是双向循环网络模型的一个运行模式示例。

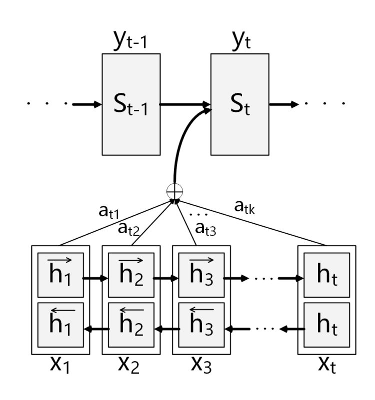


图 5 “ “症状-中药” ” 双向 RNN

掩盖（mask）和覆盖（coverage）机制

本文结合医生诊断的思考过程来分析，会形成这样一种思路，之前已经生成的中药在生成下一种中药的时候绝对不会再考虑，已经考虑过并且已经针对性的开出了中药的症状，在接下来的思考中必然不会再优先考虑。类似思路在自然语言处理里面的统计翻译方向里很早就有应用，近些年来随着深度学习的发展，很多研究者将这种思路应用到了神经网络之中，比如李航等人提出的覆盖机制[124]。北京大学的李炜[132]等人结合了李航的方法，将其应用到了中医处方生成模型上。

本文首先看输出中药的方面，中医处方和机器翻译的一个明显不同就是，在机器翻译出来的一个句子里，同一个词是可以出现多次的，但是在中药处方里面，同一种药是不能出现许多次的。所以一个最简单的想法就是直接将已经生成的中药的概率永远设为 0。这个想法虽然很粗暴但是实际的效果非常不错，本文将这种机制成为掩盖（mask）它在每次生成一种中药之后，用掩码永远地将这种中药概率设为 0。其次是在输入症状的方面，在机器翻译的过程中，会存在对某些输入的词过度翻译的情况，就是对同一个词生成了一个翻译之后又生成了一个翻译，就原始的编码解码模型来说，很难做到一种“症状-中药”一一对应的情况，反映在处方生成问题上就是，对于同一种症状，反复生成针对性的中药，因为治疗某一种症状的中药可能不止一种，相当于过度治疗。与此相对的是欠翻译的情况，就是某些词始终没有翻译，某些症状始终没有开出对应的药。这种过翻译和欠翻译的情况对于生成处方的质量产生了十分巨大的影响，无法满足临床诊断需求。

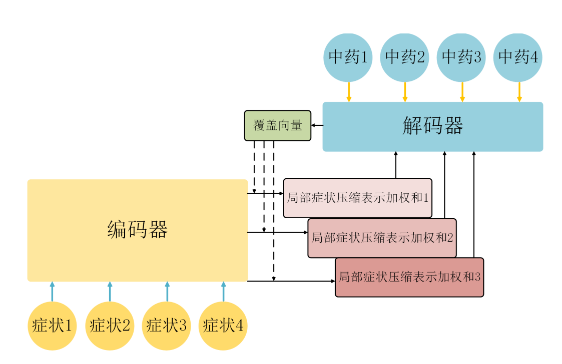


图 6 “症状-中药”结合覆盖机制与注意力机制的~~自~~编码器模型

针对这种问题，在早期的统计翻译研究里面，研究人员给每个词添加了一个计数器，来记录其翻译情况。在现在的深度学习方法中，许多研究者给出了更加精细的解决方法。 李航等[124]在一般模型里添加了一种覆盖（coverage）向量，将其与前面提到的注意力（attention）机制结合。如图 6 所示，其具体是在上文提到的局部症状压缩表示的加权和上再加一个修饰向量，这个修饰向量会根据当前已经考虑过的症状情况，适当增加或降低某个局部症状压缩表示的权重。本文最终解码某种中药所应用的是前面所有局部症状压缩表示的加权和，其权重就是注意力机制所要计算的代表每个症状重视程度的系数。而本文的覆盖向量，就是结合之前生成中药情况，比如本文之前生成的中药是针对某些症状的，那么覆盖向量就会把这些症状的局部压缩权重调低，反之就调高。在实际操作中，研究者主要提出了两个具体实现，第一个是从具体的语言学角度，具有可解释性，但是比较复杂。第二个是从神经网络的角度，比较易于实现，在李炜等人模型里采用第二种。

4 基于 transformer 的肺癌处方智能生成

2017 年，谷歌提出 transformer 模型[143]，该模型直接采用本文之前提到的 attention 模型来作为编码器-解码器的基本组件。谷歌在一系列都后续研究中，利用这篇论文提出的模型刷新了很多自然语言处理任务的记录，为提升模型的性能，本文引入 transformer 代替传统 RNN 作为编码模块。

本文构建该模型的基本思路是利用注意力机制 attention 模型完全替代了循环神经网络的方法。解决了 RNN 实际运行中的缺陷，第一，循环神经网络的一个很严重问题是无法并行化，输入必须一个接着一个，然而 attention 把串行的输入，变成了一次性的几个矩阵相乘的运算，非常适合目前 GPU 性能大幅提升的硬件进步背景。第二，传统循环神经网络模型有一个明显的缺点就是对于输入的顺序很敏感，这在翻译问题中是必要的，但是在中医处方生成方面就显得画蛇添足了，根据本文的测试，同样的一组症状输入进去，如果输入的顺序与训练模型时症状输入的顺序有明显的差异，那么输出的中药处方质量会很差。在中医临床诊断过程中，处方的症状都是医生根据个人习惯书写出来的，如果模型对处方中词语顺序太敏感的话，进行实际临床应用推广会受到很大限制。针对以上问题本文结合 transformer 模型将症状种类和顺序分开用两个模块来考虑，通过修改测试，尝试去除处理输入症状顺序的模块，再将输入测试的症状给打乱，发现乱序和正序效果并没有明显差别，这个结果就说明 transformer 模型比基于循环神经网络的编码器-解码器模型更适用于中医处方的生成。第三，因为 transformer 模型对于编码完全就是几个矩阵相乘的运算，其中的中间产物就是本文之前说的类似于局部向量压缩表示的权重系数，本文可以从程序里面讲这个权重系数矩阵提取出来，可以随时的看到在生成哪些药的时候，哪些症状被着重考虑了。

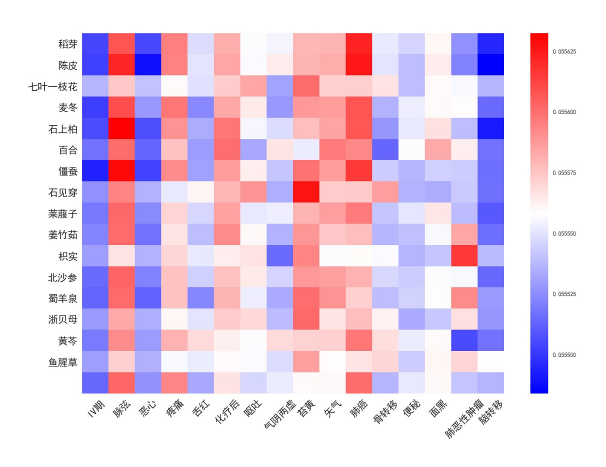


图 7 “症状-药物”关系热图

第三，因为 transformer 模型对于编码完全就是几个矩阵相乘的运算，其中的中间产物就是本文之前说的类似于局部向量压缩表示的权重系数，本文可以从程序里面讲这个权重系数矩阵提取出来，可以随时的看到在生成哪些药的时候，哪些症状被着重考虑了。简单来说，就是 transformer 模型对于中医处方的生成任务有着很好的解释性。这在目前的深度学习科研里面是很重要的，因为很多深

度学习模型最大的问题就是不知道程序里面是怎么起作用的，仅仅是知其然而不是知其所以然。图 6 是利用基于 transformer 的处方生成模型运行过程中提取出来的某个注意力权重系数矩阵画出的“症状-中药””热图，图 7 实例说明了基于 transformer 的处方生成模型方法能一定程度的准确反映实际的““症状-中药””关系，提升了模型性能。同时，我们分别利用基于 RNN 的处方生成模型和基于 transformer 的处方生成模型生成肺癌处方，进行专家评估。本文邀请了上海中医药大学附属龙华医院两位专门从事肺癌中医诊治的主治医师来对模型结果打分。上海市龙华医院肿瘤科是全国最早开阵中医药和中西医结合治疗肿瘤的专科之一，是国家中医临床研究基地（恶性肿瘤），创建以来在国医大师刘嘉湘教授带领下，在国内首倡扶正法治疗恶性肿瘤，总结了一套行之有效的中医药治疗恶性肿瘤的中医诊疗体系[154]，两位医师均有 10 年以上的临床经验，请他们针对智能处方系统生成的处方进行打分，分数在 0-10 分之间，与模型性能参数对比不同，专家的评估在一定程度上更合理，更贴近临床，两位位专家最终对基于 RNN 的处方生成模型评分为 7 分，对基于 transformer 的处方生成模型评分为 8.5 分，从专家评判的结果来看，基于 transformer 的处方生成模型生成的处方更合理。专家均认为本文构建的智能处方生成模型对临床十分有意义，在目前所达到的阶段其生成的处方可以作为门诊患者诊疗后的基础处方，由专业医师在其基础上根据实际情况再进行调整，这样可以大大提高门诊医师的诊疗效率。

5 总结

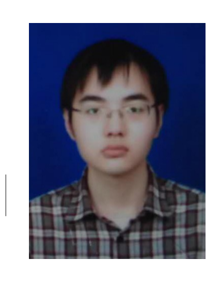
与之前基于机器学习的不区分病种的中医大数据分析不同。本文从提升中医临床诊断能力和创新发展需求出发，以肺癌诊断为实际临床场景，探索人工智能在临床和科研中的实际运用。本文提出基于深度学习的智能肺癌处方模型。本文巧妙地将中医临床处方的生成建模为一个是一个机器翻译的问题，利用基于双向循环神经网络的自编码器输入症状输出中药组合，并进一步使用 transformer 模型替代 RNN 作为自编码模块提升模型性能。根据处方数据中药物与症状关系等特征，引入注意力机制和覆盖机制提升模型的准确性和鲁棒性。未来工作，将更多中医知知识，例如药物用量以及病证等加入生成模型并引入已生存期为疗效指标，从海量中医临床处方中挖掘最佳的诊疗模式，建立智慧中医全周期治疗模型。



阮春阳

复旦大学计算机科学与技术学院博士生。主要研究方向：智慧中医，图表示学习，图神经网络。

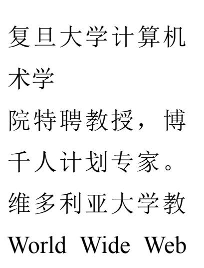
cyruan16@fudan.edu.cn



裴朝翰

复旦大学计算机科学与技术学院硕士生。主要研究方向：智慧中医，文本挖掘，深度学习。

17212010027@fudan.edu.cn



张彦春

复旦大学计算机科学与技术学院特聘教授，博导，国家千人计划专家。澳大利亚维多利亚大学教授；World Wide Web 期刊主编 Health Information Science and Systems 期刊主编。主要研究方向：数据挖掘、医疗与健康数据分析 。

YanchunZhang@fudan.edu.cn



杨蕴

上海中医药大学附属龙华医院在读博士，上海中医药大学附属上海市中西医结合医院肿瘤科主治医师。上海市抗癌协会传统医学专业委员会青年委员。主要从事恶性肿瘤中医药治疗的临床和基础研究。



田建辉

主任医师，医学博士，博士生导师。上海中医药大学附属龙华医院肿瘤科主任医师，上海市中医药研究院中医肿瘤研究所基础研究部主任。中国中医药学会中医肿瘤分会常务委员，美国癌症研究协会（AACR）协会活跃会员。

参考文献

[1] 张晓航,石清磊,王斌,王炳蔚,王永吉,陈力,吴敬征。机器学习算法在中医诊疗中的研究综述[J].计算机科学,2018,45(S2):32-36.
[2] Huimin Luo, Min Li, Shaokai Wang, Quan Liu, Yaohang Li, Jianxin Wang,Computational drug repositioning using low-rank matrix approximation and randomized algorithms. Bioinformatics 34(11): 1904-1912 (2018).
[3] Anahita Hosseini, Ting Chen, Wenjun Wu, Yizhou Sun, Majid Sarrafzadeh,HeteroMed: Heterogeneous Information Network for Medical Diagnosis. CIKM 2018: 763-772.
[4] Xiaofang Zhou, Xue Li, Yangyang Hu, Wenqiang Zhang, Fufeng Li,Lip analysis in traditional Chinese medicine. BIBM 2017: 1381-1387.
[5] 刘嘉湘. 发挥中西医优势治疗肺癌[J]. 中医学报,2014(03):314-315.
[6] 张德政，哈爽，刘欣，等. 中医药领域人工智能的研究与发展[J]. 情报工程,2018,4(01):13-2.
[7] 郭玉萱,阮春阳,王晔,张彦春.基于不平衡数据分类的中药肝毒性检测[J].计算机应用与软件,2018,35(08):226-230.
[8] Xintian Chen, Chunyang Ruan, Yanchun Zhang, Huijuan Chen:Heterogeneous Information Network Based Clustering for Categorizations of Traditional Chinese Medicine Formula. 839-846
[9] Chunyang Ruan,Jiangang Ma, Ye Wang, Yanchun Zahng, Yun Yang,Discovering Regularities from Traditional Chinese Medicine Prescriptions via Bipartite Embedding Model.IJCAI 2019.
[10] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.
[11] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409 。0473, 2014.
[12] Tu Z, Lu Z, Liu Y, et al. Modeling coverage for neural machine translation[J]. arXiv preprint arXiv:1601.。04811, 2016.
[13] Li W, Yang Z, Sun X. Exploration on generating traditional chinese medicine prescription from symptoms with an end-to-end method[J]. arXiv preprint arXiv:1801.09030, 2018.
[14] Vaswani A, Shazeer N, Parmar N, et al。Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
[15] 上海中医药大学附属龙华医院肿瘤科国家中医临床研究基地(恶性肿瘤)[J].上海中医药杂志,2010,03(03):86.

英文题目：Wise Traditional Chinese Medicine: Intelligent Formulation Model of Lung Cancer Prescription

英文摘要：

Traditional Chinese medicine (TCM) plays an important role in the comprehensive treatment system for lung cancer. The quality of TCM prescription depends on doctors' clinical knowledge and technical level. We input the handwritten clinical Chinese medicine prescription, which is regarded as the source data, into the improved encoder and translation model in order to find the corresponding relationship between Chinese medicine and symptoms. And then, we use these rules to automatically generate effective lung cancer prescriptions.